ED 457 195 TM 033 293

AUTHOR	Capar, Nilufer K.
TITLE	Analyzing Multidimensional Response Data Structure Represented by Unidimensional IRT Models To Increase the Precision of Scoring Using Out-of-Scale Information.
PUB DATE	2000-11-00
NOTE	38p.
AVAILABLE FROM	Paper presented at the Annual Meeting of the Florida Educational Research Association (45th Tallahassee, FL, November 8-10, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Ability; *Adaptive Testing; *Computer Assisted Testing; *Item Response Theory; Mathematics; Reading Comprehension; Scaling; *Scores; Simulation; *Test Bias; Test Items

ABSTRACT

This study investigated specific conditions under which out-of-scale information improves measurement precision and the factors that influence the degree of reliability gains and the amount of bias induced in the reported scores when out-of-scale information is used. In-scale information is information that an item provides for a composite trait to which it is specifically classified by a content review. Out-of-scale information is defined as information that an item provides for a composite trait other than the composite to which it is specifically classified. Computer simulated two-dimensional data were used to investigate the effects of various factors in the precision of reported scores computed with and without out-of-scale information. Different testing conditions were specified to compare the performance of the traditional information computation method (in-scale information alone; Method A) and the alternative method (in- and out-of-scale information together; Method B). The findings for standard error indicate that random measurement error is more evident with Method A. Bias results show that ability estimates at high and low theta levels are pulled toward the mean with both methods. The root mean square errors reflecting the combined effects of standard errors and bias differed with different conditions and test lengths, as discussed. The choice between the in-scale and out-of-scale procedures would seem to depend primarily on standard error and bias. Because Method B resulted in lower standard errors and usually performed better with the short test moderate discrimination conditions, this alternative method is likely to be attractive. Using out-of-scale information is likely to produce reliable subscores that can be used for diagnostic purposes. (Contains 4 tables, 11 figures, and 25 references.) (SLD)

Analyzing Multidimensional Response Data Structure Represented by Unidimensional IRT Models to Increase the Precision of Scoring Using Out-of-Scale Information

NILUFER K. CAPAR

Florida State University

Paper presented at the Annual Meeting of the Florida Educational Research Association, Tallahassee, Florida, November 2000

Background

The unidimensional item response theory (IRT) models have the advantage of being mathematically fairly simple. Yet, it is likely that the actual interactions between examinees and test items are not as simple as is implied by the unidimensional IRT models. The multidimensional nature of tests has been a concern to measurement professionals since early eighties (e.g., Ansley & Forsyth, 1985). Examinees are likely to rely on more than a single ability answering a particular test question, and sometimes the test questions require a combination of a number of skills or abilities to determine a correct solution. Multidimensional IRT (MIRT) applications are not commonly used in modeling and scoring test data, primarily because MIRT applications require a great deal of item response data to adequately estimate item and person parameters. However, using unidimensional IRT models sacrifices information about the trait levels of examinees (Miller & Hirsh, 1992).

There are two general approaches in the literature that have proposed making better use of information provided by all test items, which is not used in conventional IRT applications, to increase subscore reliabilities: test level and item level information augmentation.

The test score level approach uses an empirical Bayes procedure in combination with IRT to adjust unstable subscores. Yen (1987) developed "objective performance index" (OPI) that can be used to provide reliable subscores with diagnostic value. An examinee's overall test performance is used as a priori to adjust subscale scores. Wainer, Sheehan, and Wang (2000) presented a similar procedure that differed from the Yen's procedure by assuming a normal distribution instead of a binominal distribution for the total test score and allowing subscores to be used instead of the total test score in stabilizing subscale scores. The subscale score, however, is not independent from the total test score with the Bayesian procedures.

The item level approach is based on fitting IRT models to response data with reference to a composite trait at a time, which is defined by the items composing a subscale. Luecht and Miller (1992) demonstrated that it was possible to retain multidimensional interpretations of composite traits, even though test calibrations are conveniently unidimensional. The common problem of assuming unidimensionality for an entire test is avoided by restricting that assumption to items shown to directionally cluster together in multidimensional space. Ackerman and Davey (1991) illustrated how collateral information, information that items provide for a correlated trait, to improve ability estimates in Computer Adaptive Testing (CAT). Davey and Hirsh (1991) designed a simulation study to compare concurrent and consecutive parameter estimation procedures. The conventional parameter estimates were referred as to concurrent parameter estimates, and the proposed parameter estimates referred as to consecutive parameter estimates. Consecutive parameter estimation involved calibrating items with respect to both a primary trait and a non-primary trait. Associated estimation bias and standard error estimates were reported as well as the performance of consecutive parameter estimates in identifying unusual examinees.

Both approaches are reported to increase subscore reliabilities with the expense of some bias: extreme scores were pooled toward the average ability score. The bias component in the out-of-scale information is independent from the ability domain measured by a subscale, unlike the OPI procedure where the total test score is dependent on the ability domain measured by the subscale. The authors, however, agreed on that the bias involved in the estimates is compensated with the increase observed in the reliabilities of subscores that might not be reported otherwise.

This study purposed to extent on current literature and investigated specific conditions under which out-of-scale information improves measurement precision, and to determine the factors that influence both the degree of reliability gains and the amount of bias induced in the reported scores when out of scale information is used. Proposed methodology in this study provides means for analyzing the expected contribution of out-of-scale information that can be used to increase the number of scores for diagnostic purposes without increasing the number of items or testing. This study used the composite traits approach that allows obtaining item parameter estimates for an item on a non-primary but related trait composite by projecting that item as a vector onto the non-primary trait composite. It was expected that the higher the correlation between two score composites the greater would be the benefit from using out-of-scale information.

The expected increase in bias is a potential drawback using out-of-scale information to increase precision of test scores. The bias, the difference between estimated and true ability estimates, is expected to increase some degree due to included secondary information source (Davey & Hirsh, 1991). The ultimate aim is to delineate conditions under which the hypothesized increase in the information provided for examinees compensates for the expected bias introduced to the scoring procedure.

Geometrical Representation of In- and Out-of-Scale Parameters

In most cases, items are classified to measure a single trait or trait composite even though they measure more than one content domain or trait at a time. providing some bonus information. If this bonus information is with reference to a trait that we want to report scores, it can be incorporated into the scoring procedure, and may result reliable scores that may not be reported otherwise. For example, science items may have a strong math component. We can use the available information in the response data by letting these items contribute to the science score as well as the math score. Furthermore, items that discriminate among examinees in a narrow content domain can be used to stabilize subscores for other narrow content domains within a single test. For example, Algebra items in a math test can be used to increase the reliability of the Geometry subscale score as long as those math items discriminate among examinees in the measured Geometry domain.

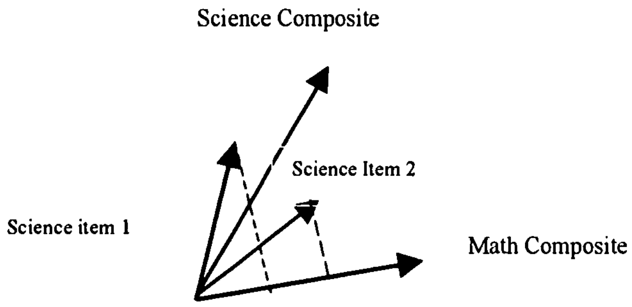


Figure 1 Projections item vectors onto composites being measured

The solid arrows in Figure 1 represents two science items as vectors in a two-dimensional space formed by the science and math. The direction of the item vector indicates the direction in space the item best measures, while the length of the vector indicates how discriminating the item is in that direction of the space. Interpreted geometrically, an item discriminates with respect to a composite proportionally to the length of its projection on that composite. The dashed lines in the figure show how the discriminations of the two science items diminish when they are projected onto the math composite.

Methods

As mentioned earlier, two types of information that an item can provide are distinguished: In-scale and out-of-scale. In-scale information is defined as information that an item provides for a composite trait to which it is specifically classified by a content review. Out-of-scale information is defined as information that an item provides for a composite trait other than the composite to which it is specifically classified. Computer simulated two-dimensional data was used to investigate the effects of various factors on the precision of reported scores computed with and without out-of-scale information. A variety of testing conditions were specified to compare the performance of the traditional information computation method, (in-scale information alone) and the proposed alternative information computation method (in- and out-of-scale information together). Factors investigated were (a) test length, (b) the magnitude of item discrimination, and (c) the degree of association between pairs of traits.

Evaluation criteria were information provided and indices of estimation errors: standard error (SE), bias, and root mean squared error (RMSE). These criteria were chosen to evaluate the precision of reported ability estimates or scores that may lead one to choose the in- and out-of-scale information computation method over the in-scale information computation method in practical assessment situations. Factors and evaluation criteria used are described below in detail.

Information Computation Methods

Two information computation methods are defined: Method A and Method B. Method A uses information provided by only in-scale items, while Method B uses information provided by both in- and out-of-scale items in computing evaluation criteria indices for examinees at their true ability levels. It is hypothesized that the information provided for examinees would increase when Method B is used. It is also hypothesized that information increase would become greater with Method B when there is a stronger association between the in- and out-of-scale traits.

Modeling Multidimensional Response Data

A test battery was simulated to model two-dimensional (2-D) response data for examinees from a multivariate normal distribution. The battery consisted of two tests each measuring a single trait (or a trait composite). The proposed procedure is expected to be more beneficial when data at hand include examinee responses to more than two tests, allowing simultaneous inferences to be made. However, the challenge faced is in the complexity of controlling response data when the number of dimensions is more than two. Response data for a two-test battery was modeled in this study. The response data included examinees' responses to two content domain scales. The two-test battery composite measures are given generic names for convenience and to as math and science.

Factors to be investigated

Two factors to be kept constant in this study were examinee sample size as 1000 and multidimensional structure of the response data as two-dimensional across all conditions.

Test Length

Two test battery lengths are included: a 30-item battery composed of two 15-item subtests and a 60-item battery composed of two 30-item subtests. Test length was varied to evaluate the relationship between the hypothesized increase in the information provided and the number of items included in the battery.

Item Discrimination Power

Item discrimination parameters were varied in generating the response data for the in-scale three-parameter logistic (3PL) item parameters of the fitted IRT model resulting in a high-discrimination ideal test battery and a moderate-discriminating test battery. It was expected that the precision increase would be greater for the high-discrimination condition, since information provided by an item is primarily a function of its discrimination parameter.

The overall means of the difficulty, discrimination, and pseudo-guessing parameter values were selected to be 0.00, 1.00, and 0.15, respectively for the moderate-discrimination condition. The item discrimination parameters of the moderate-discrimination battery were drawn from a normal distribution with a mean of 1.00 and standard deviation (SD) of 0.10. Only the overall mean of the item discrimination parameter value was changed for the high-discrimination condition and set as 1.80. The item discrimination parameters of the high-discrimination battery were drawn from a normal distribution with a mean of 1.80 and standard deviation (SD) of 0.10. The value 1.80 was selected to be the mean of item discrimination parameter in the high-discriminating test battery.

Table 1.Summary Statistics for the Item Parameters of the 30-Item (Short) Test-Battery

Battery/Parameter	Number of Items	Mean	SD	Min.	Max
Moderate-discriminating test battery					
Math					
a[a]	15	0.9759	0.0936	0.7807	1.1086
b[b]	15	0.0004	1.5974	-2.5008	2.5000
c[c]	15	0.1500	0.0000	0.1500	0.1500
Science					
a	15	0.9667	0.1058	0.7804	1.1010
b	15	0.0004	1.5974	-2.5008	2.5000
c	15	0.1500	0.0000	0.1500	0.1500
High-discriminating test battery					
Math					
a	15	1.7856	0.0974	1.5852	1.9491
b	15	0.0004	1.5974	-2.5008	2.5000
c	15	0.1500	0.0000	0.1500	0.1500
Science					
a	15	1.7652	0.0923	1.5785	1.9598
b	15	0.0004	1.5974	-2.5008	2.5000
c	15	0.1500	0.0000	0.1500	0.1500

Table 2.Summary Statistics for the Item Parameters of the 60-Item (Long) Test-Battery

Battery/Parameter	Number of Items	Mean	SD	Min.	Max
Moderate-discriminating test battery					
Math					
a[a]	30	1.0281	0.1322	0.8447	1.2659
b[b]	30	-0.0012	1.5186	-2.5025	2.5000
c[c]	30	0.1500	0.0000	0.1500	0.1500
Science					
a	30	0.9958	0.1069	0.8662	1.1806
b	30	-0.0012	1.5186	-2.5025	2.5000
c	30	0.1500	0.0000	0.1500	0.1500
High-discriminating test battery					
Math					
a	30	1.7919	0.0967	1.6295	1.9987
b	30	-0.0012	1.5186	-2.5025	2.5000
c	30	0.1500	0.0000	0.1500	0.1500
Science					
a	30	1.7969	0.1208	1.5973	2.0243
b	30	-0.0012	1.5186	-2.5025	2.5000
c	30	0.1500	0.0000	0.1500	0.1500

Note. The pseudo-guessing parameter values were fixed at 0.15

[a] Item discrimination parameter value

[b] Item difficulty parameter value

[c] Item pseudo-guessing parameter value

The item difficulty parameters for both conditions were equally spaced values from the interval between –2.5 and 2.5. Summary statistics of the item parameters generated in the short test length and the long test length condition are summarized in Table 1 and Table 2 for both moderate- and high-discrimination test batteries.

Trait Composite Correlations

It was hypothesized that the increment of information in Method B, the in- and out-of-scale information computation method would increase as the correlation between the two traits involved increase. This hypothesis was evaluated by varying the correlations between two trait composites in generating response data forming five conditions: 0.10, 0.30, 0.50, 0.70, and 0.90.

Simulation

The RESGEN computer program (Muraki, 1992) was used to generate 2-D test battery response data for 1000 examinees sampled from a multivariate normal distribution. Response data were generated for a total of twenty conditions (2 test length X 2 item discrimination X 5 correlation conditions). The multidimensionality in response data was modeled as a complex two-dimensional structure where each item had two discriminating parameters, one item discrimination parameter for each trait measured.

True ability parameters and true item parameters were fixed across replications, and a randomly selected seed number was used for each replication.

Item Calibration

In-scale and out-of-scale item parameters were estimated using the BILOG (Mislevy & Bock, 1990) and the PIC (Davey & Spray, 1999) computer programs fitting the 3PL IRT model to each test. The probability of a correct response in the 3PL model is given by

$$P_i(\theta)=c_i+(1-c_i)\left[1+e^{-Da_i(\theta-b_i)}\right]^{-1}, \qquad (1)$$

where $_i$ is the item administered, a_i is the discrimination, b_i is the difficulty, and c_i is the pseudo-guessing parameter of item $_i$, D is the scaling constant 1.702 and P_i (•) is the 3PL model probability of a correct answer for the ith item for an examinee with ability θ (Hambleton & Swaminathan, 1985).

PIC uses the method of maximum likelihood to calibrate out-of-scale items one at a time for a composite trait. The in-scale item parameters are held constant in every PIC run to fix the scale, and out-of-scale items are calibrated individually, as the presence of other out-of-scale items would contaminate the scale. For example, out-of-scale parameters of math items were calibrated one at a time with respect to the Science composite, which was defined by all the items in the science test, holding estimated science in-scale item parameters constant. Each item in the two-test battery had two sets of parameters: one in-scale and one out-of-scale.

Evaluation Criteria

Indices of estimation errors (RMSE, SE and bias) were calculated for each examinee at his/her true ability using both methods. The average Information provided for each simulee was computed over 20 replications under each condition. Descriptive statistics were provided for 24 ability intervals on true ability or theta scale with the increments of 0.2.

Information

Two information functions were computed for each item: one for Method A and one for Method B by

$$I(\theta_i) = \frac{(p_i^{`})^2}{P_i Q_i}, \tag{2}$$

where P_i is the 3PL model probability of a correct answer for the ith item for jth examinee with ability θ, $P`_i$ is the second derivative of P_i. Equation 2 is estimated by

$$I(\theta_i) = D^2 a_i^2 P_i Q_i \left[\frac{P_i^*}{P_i} \right]^2, \tag{3}$$

where,

$$P_i^*(\theta) = \frac{P_i(\theta) - c_i}{1 - c_i}. \tag{4}$$

For example, each math item had two information functions in the two-test battery, one with respect to the math composite, in-scale, and one with respect to the science composite, out-of-scale.

Total information provided for jth examinee with ability θ_j was computed by

$$I(\theta_j) = \sum_{i=1}^{n} \left(\frac{(P_i')^2}{P_i Q_i} \right), \tag{5}$$

where $_i$ is the item administered, P_i is the three-parameter logistic (3PL) model probability of a correct answer for the ith item for jth examinee with ability θ, P'_i is the second derivative of P_i, and $I(\theta_j)$ is information provided for the jth examinee at its true ability by summing the information provided by $i=(1,\ldots,n)$ items (Hambleton & Swaminathan,1985).

In method A, the total information were computed for each examinee's true math and science ability levels by summing the information provided by in-scale items in the test battery. In method B, the total information were computed for each examinee's true math and science ability levels by summing the information provided by all (in- and out-of-scale items) the items in the test battery. There were 15 items in the short test-battery and 30 items in the long test-battery to be summed over for Method A. While, there were 30 items in the short test-battery and 60 items in the long test-battery to be summed over for Method B.

Analysis

Ability estimates and information provided for examinees, and indices of ability estimation errors were compared for Method A and Method B along the ability scale. The stability of the study results across conditions were assessed across conditions for replications, which are 20 in this study. The root mean square error (RMSE) were computed across conditions for the ability estimates obtained with (Method B) and without (Method A) out-of-scale information. Bias, SE, and RMSE were computed for each examinee using these formulas:

$$Bias\left(\theta_j\right)=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_j-\theta_j\right) \tag{6}$$

$$SE\left(\theta_j\right)=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_j-\frac{\sum_{j=1}^{N}\hat{\theta}_j}{N}\right)}, \tag{7}$$

$$RMSE\left(\theta_j\right)=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_j-\theta_j\right)^2}, \tag{8}$$

where, θ is the true ability, estimatedθ_j is the ability estimate for the jth replication, and N is the number of replications, which is 20 in this study. The RMSE^2 is an index of parameter recovery composed of two parts, one reflecting the variance (SE^2) of the estimates and the other reflecting the bias (BIAS^2) in estimation (Gifford & Swaminathan, 1990; Wang & Vispoel, 1999). When the estimated SE is small it suggests that random error in ability estimates are small. When the estimated bias is small it suggests that systematic error in ability estimates are small. Larger RMSE values for any condition indicate that the procedure used is working poorly in terms of recovering the true parameters over replications.

Results

Table 3 lists the conditions used. The mean conditional plots were provided for the math subtest only, which were very similar to those of the science test as expected since only the randomly selected item parameters were not identical for both tests.

Estimating Ability

BILOG is used to obtain ability estimates for simulees for both methods. As mentioned in earlier, unidimensional in-scale and out-of-scale item parameters were estimated by the maximum likelihood estimator using the BILOG (Mislevy & Bock, 1990) and the PIC (Davey & Spray, 1999) computer programs fitting the 3PL IRT model to each test.

In method A, BILOG was used to calibrate items (in-scale) and to obtain ability estimates. In Method B, BILOG was provided a parameter file that included both BILOG calibrated in-scale item parameters and PIC calibrated out-of-scale items to obtain ability estimates. Even though any of the available estimation procedures could be used with BILOG, the expected a priori, (EAP) procedure was used to obtain ability estimates for conditions studied.

Although the maximum likelihood, (ML), estimator is an unbiased estimate of the sample standard deviation, unlike the EAP estimator, a well-known problem arises with the use of the ML estimator when there are extreme response patterns, (i.e., all correct or incorrect responses), that produce extreme ability estimates. For short tests, tests with less than 20 items, the probability of observing such problematic response patterns increases. The results are very large ability estimate ranges, that is if a reasonable convergence level is reached for such response patterns. The BILOG manual recommends the use of EAP procedure to obtain ability estimates to refrain from such problematic cases when tests are composed of less than 20 items. This recommendation is followed in this study since 10 of the total tests studied in 20 conditions had two 15-item tests. The BILOG Manual states that in most applications the bias effect in EAP is not apparent if the prior distributions are selected to be the same and ability distributions are put on a common scale. Ability estimates in the current study were obtained using EAP Bayesian procedure using a normal prior distribution with a mean of zero and variance of 1. All ability estimates were rescaled using BILOG with a mean of zero and variance of 1 for all conditions.

Table 3. Study Conditions and Factors

Math and Science Test Battery	**Test length**				**Average Item Discrimination for Both Tests**	**Math and Science Trait Correlation**
	Method A		*Method B*			
Conditions	Math	Science	Math	Science		
Condition 1	15	15	30	30	Moderate	0.1
Condition 2	15	15	30	30	Moderate	0.3
Condition 3	15	15	30	30	Moderate	0.5
Condition 4	15	15	30	30	Moderate	0.7
Condition 5	15	15	30	30	Moderate	0.9
Condition 6	15	15	30	30	High	0.1
Condition 7	15	15	30	30	High	0.3
Condition 8	15	15	30	30	High	0.5
Condition 9	15	15	30	30	High	0.7
Condition 10	15	15	30	30	High	0.9
Condition 11	30	30	60	60	Moderate	0.1
Condition 12	30	30	60	60	Moderate	0.3
Condition 13	30	30	60	60	Moderate	0.5
Condition 14	30	30	60	60	Moderate	0.7
Condition 15	30	30	60	60	Moderate	0.9
Condition 16	30	30	60	60	High	0.1
Condition 17	30	30	60	60	High	0.3
Condition 18	30	30	60	60	High	0.5
Condition 19	30	30	60	60	High	0.7
Condition 20	30	30	60	60	High	0.9

Note: Test data were two dimensional for the 20 conditions.

Information Provided

Table 4 shows the average information and the average estimated reliability for each test using Method A and Method B under each condition.

The reliabilities of both subtest scores invariably increased when Method B (in- and out-of-scale information procedure) was used. The reliability gain became greater as the correlation between the math and science composites increased from 0.5 to 0.7 for both moderate and high discrimination. Even though the increase was greater for moderately high to high correlation conditions, the scales that most benefited from the gain were the short test-moderate discrimination conditions and the long test-moderate discrimination conditions.

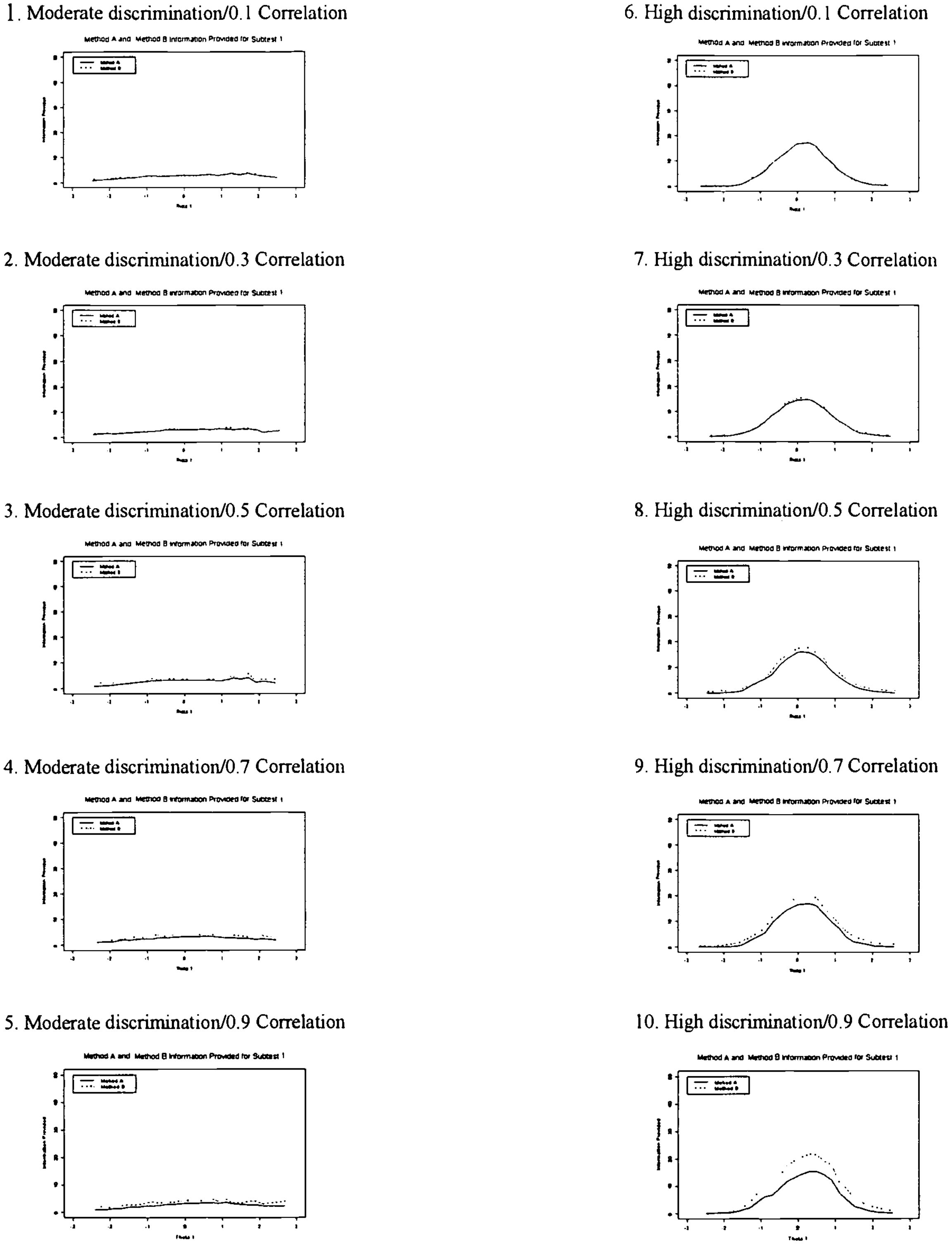
The reliability estimates of the short test conditions increased from approximately 0.65 to 0.72 when Method B was used for moderate discrimination (conditions 1-5), and from approximately 0.89 to 0.90 for high discrimination conditions. The increase was capitalized when the correlation between the math and science composites was 0.5 or greater (conditions 6-10).

The reliability estimates of the long test conditions increased from 0.85s to 0.89s for moderate discrimination conditions, and from 0.95s to 0.97s for high discrimination conditions when the correlation between the math and science composites was 0.5 or greater (conditions 13-15 and 18-20). However, the gain observed seems to be of practical use for short test-high discrimination and long test-moderate discrimination conditions allowing an argument to be made whether the associated subscale scores are reliable enough to report.

Table 4. Average Information and Reliabilities over 20 Replications

Condition		MATH TEST						SCIENCE TEST					
		TRADITIONAL METHOD			ALTERNATIVE METHOD			TRADITIONAL METHOD			ALTERNATIVE METHOD		
	Statistic	INFO	SE	REL	INFO	SE	REL	INFO	SE	REL	INFO	SE	REL
1	Mean	2.780	0.360	0.640	2.898	0.346	0.654	2.697	0.371	0.629	2.822	0.355	0.645
	Var	0.014	0.000	0.000	0.015	0.000	0.000	0.015	0.000	0.000	0.014	0.000	0.000
2	Mean	2.8498	0.352	0.648	3.077	0.326	0.674	2.667	0.376	0.624	2.908	0.345	0.655
	Var	0.020	0.000	0.000	0.024	0.000	0.000	0.017	0.000	0.000	0.020	0.000	0.000
3	Mean	2.818	0.355	0.645	3.495	0.287	0.713	2.683	0.374	0.627	3.370	0.297	0.703
	Var	0.013	0.000	0.000	0.021	0.000	0.000	0.015	0.000	0.000	0.022	0.000	0.000
4	Mean	2.818	0.355	0.645	3.495	0.287	0.713	2.683	0.374	0.627	3.370	0.297	0.703
	Var	0.013	0.000	0.000	0.021	0.000	0.000	0.015	0.000	0.000	0.022	0.000	0.000
5	Mean	2.847	0.352	0.648	3.904	0.257	0.743	2.751	0.364	0.636	3.822	0.262	0.738
	Var	0.022	0.000	0.000	0.031	0.000	0.000	0.015	0.000	0.000	0.022	0.000	0.000
6	Mean	9.876	0.101	0.899	9.984	0.100	0.900	9.219	0.109	0.891	9.334	0.107	0.893
	Var	0.103	0.000	0.000	0.104	0.000	0.000	0.109	0.000	0.000	0.108	0.000	0.000
7	Mean	9.482	0.106	0.894	9.876	0.101	0.899	9.423	0.106	0.894	9.832	0.102	0.898
	Var	0.173	0.000	0.000	0.181	0.000	0.000	0.167	0.000	0.000	0.175	0.000	0.000
8	Mean	9.560	0.105	0.895	10.600	0.094	0.906	9.232	0.108	0.892	10.219	0.098	0.902
	Var	0.091	0.000	0.000	0.116	0.000	0.000	0.040	0.000	0.000	0.053	0.000	0.000
9	Mean	9.630	0.104	0.896	11.966	0.084	0.916	9.640	0.104	0.896	12.023	0.083	0.917
	Var	0.110	0.000	0.000	0.129	0.000	0.000	0.050	0.000	0.000	0.056	0.000	0.000
10	Mean	9.311	0.108	0.892	13.873	0.072	0.928	9.304	0.108	0.892	13.894	0.072	0.928
	Var	0.115	0.000	0.000	0.167	0.000	0.000	0.068	0.000	0.000	0.110	0.000	0.000
11	Mean	6.841	0.146	0.854	7.097	0.141	0.859	8.079	0.124	0.876	8.300	0.121	0.879
	Var	0.044	0.000	0.000	0.046	0.000	0.000	0.054	0.000	0.000	0.055	0.000	0.000
12	Mean	6.902	0.145	0.855	7.507	0.133	0.867	8.155	0.123	0.877	8.745	0.114	0.886
	Var	0.035	0.000	0.000	0.036	0.000	0.000	0.041	0.000	0.000	0.047	0.000	0.000
13	Mean	6.841	0.146	0.854	8.082	0.124	0.876	8.261	0.121	0.879	9.604	0.104	0.896
	Var	0.043	0.000	0.000	0.051	0.000	0.000	0.047	0.000	0.000	0.068	0.000	0.000
14	Mean	6.830	0.147	0.853	9.072	0.110	0.890	8.278	0.121	0.879	10.792	0.093	0.907
	Var	0.033	0.000	0.000	0.042	0.000	0.000	0.031	0.000	0.000	0.048	0.000	0.000
15	Mean	6.861	0.146	0.854	10.760	0.093	0.907	8.464	0.118	0.882	12.962	0.077	0.923
	Var	0.058	0.000	0.000	0.119	0.000	0.000	0.030	0.000	0.000	0.081	0.000	0.000
16	Mean	18.902	0.053	0.947	19.129	0.052	0.948	18.828	0.053	0.947	19.060	0.052	0.948
	Var	0.231	0.000	0.000	0.237	0.000	0.000	0.143	0.000	0.000	0.141	0.000	0.000
17	Mean	19.380	0.052	0.948	20.229	0.049	0.951	19.647	0.051	0.949	20.473	0.049	0.951
	Var	0.111	0.000	0.000	0.117	0.000	0.000	0.151	0.000	0.000	0.160	0.000	0.000
18	Mean	19.151	0.052	0.948	21.257	0.047	0.953	19.540	0.051	0.949	21.713	0.046	0.954
	Var	0.147	0.000	0.000	0.204	0.000	0.000	0.138	0.000	0.000	0.165	0.000	0.000
19	Mean	19.386	0.052	0.948	24.347	0.041	0.959	19.165	0.052	0.948	24.244	0.041	0.959
	Var	0.102	0.000	0.000	0.147	0.000	0.000	0.104	0.000	0.000	0.127	0.000	0.000
20	Mean	19.654	0.051	0.949	30.564	0.033	0.967	19.797	0.051	0.949	30.794	0.032	0.968
	Var	0.151	0.000	0.000	0.304	0.000	0.000	0.138	0.000	0.000	0.296	0.000	0.000

- Reliability = $\rho = (\sigma_\theta^2 - \sigma_e^2) / \sigma_\theta^2$, where $\sigma_e^2 = 1 / \int I(\theta)g(\theta)d(\theta)$, and since it is assumed that $\sigma_\theta^2 = 1$, $\rho = 1 - \sigma_e^2$
- Method A and Method B information (INFO) for the Math test were plotted in Figure 4.
- Numbers listed were rounded to three decimal places.
- Figure 4 and Figure 5 plot the first colums of the tradirional method and alternative method for the math test.



1. Moderate discrimination/0.1 Correlation

2. Moderate discrimination/0.3 Correlation

3. Moderate discrimination/0.5 Correlation

4. Moderate discrimination/0.7 Correlation

5. Moderate discrimination/0.9 Correlation

6. High discrimination/0.1 Correlation

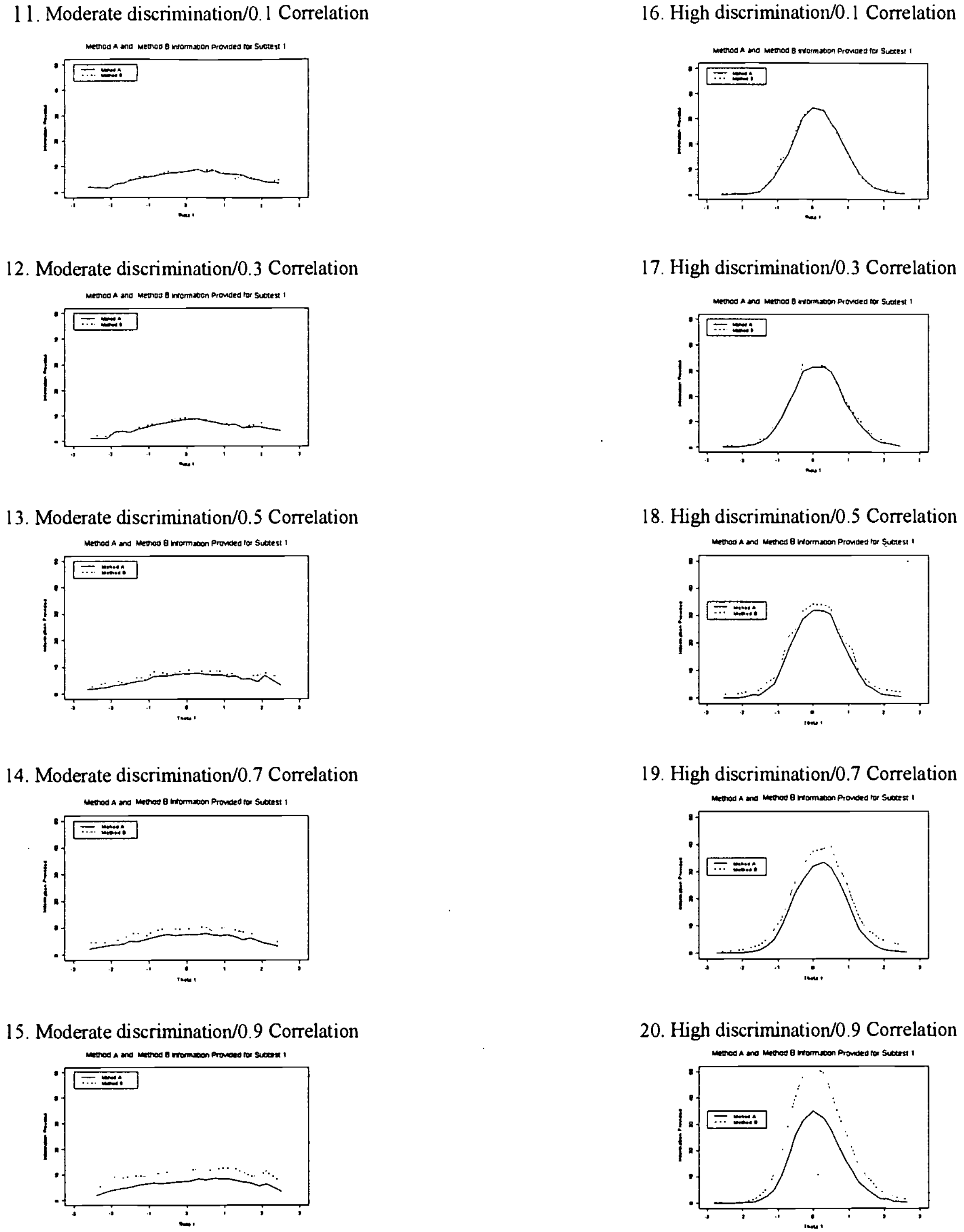
7. High discrimination/0.3 Correlation

8. High discrimination/0.5 Correlation

9. High discrimination/0.7 Correlation

10. High discrimination/0.9 Correlation

Figure 4. Information Provided for Examinees by the Math Test for the Short Test Conditions



11. Moderate discrimination/0.1 Correlation

12. Moderate discrimination/0.3 Correlation

13. Moderate discrimination/0.5 Correlation

14. Moderate discrimination/0.7 Correlation

15. Moderate discrimination/0.9 Correlation

16. High discrimination/0.1 Correlation

17. High discrimination/0.3 Correlation

18. High discrimination/0.5 Correlation

19. High discrimination/0.7 Correlation

20. High discrimination/0.9 Correlation

Figure 5. Information Provided for Examinees by the Math Test for the Long Test Conditions

Figure 4 and 5 show the conditional mean information provided for the math subtest using Method A and Method B for short test (1-10) and long test (11-20) conditions, respectively. It was hypothesized that information provided for examinees would increase with Method B, and stronger association between the in- and out-of-scale traits would contribute to this increase. The results obtained confirmed this hypothesis.

The information increase with Method B was approximately 3%, 9%, 25%, 25%, and 39% for moderate discrimination-short test conditions, and increased in magnitude as the correlation between math and science composites increased from 0.1 to 0.9. The information increase with Method B was approximately 1%, 4%, 10%, 24%, and 49% for high discrimination-short test conditions, also increased in magnitude as the correlation between math and science composites increased from 0.1 to 0.9. However, the increase was smaller for lower correlation conditions, conditions 1 through 3, than for high correlation conditions, conditions 4 and 5. A similar pattern was observed for long-test conditions, conditions 10-20, with the exception that the information gain was greater in general. Information gain observed was 3%, 7%, 17%, 30%, and 55% for the moderate discrimination-long test conditions (conditions 11-15), and 5%, 5%, 19%, 26%, and 55% for high-discrimination-long test conditions (conditions 16-20).

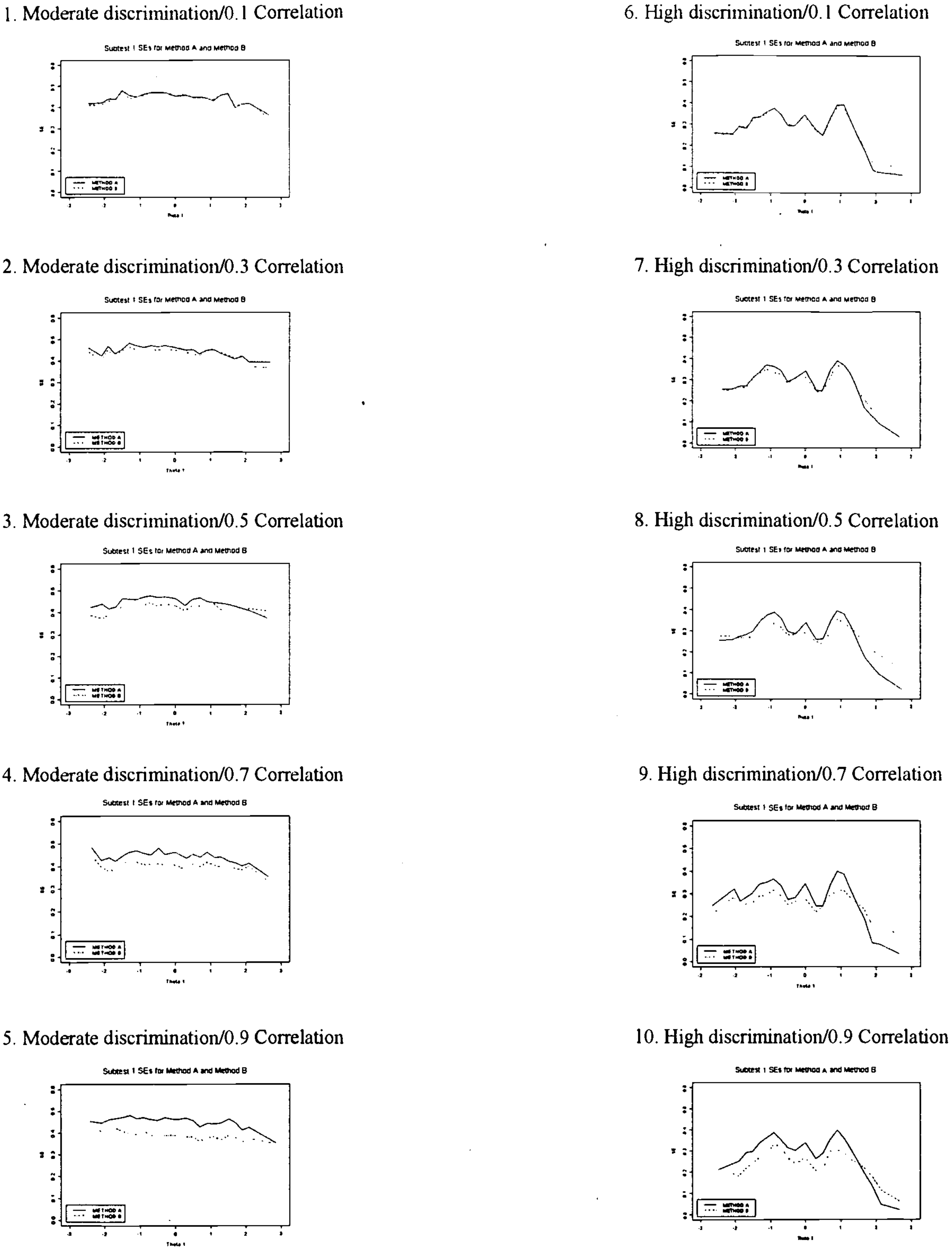
Standard Error

Figure 6 and 7 show the conditional mean standard error (SE) plots for math ability at short (1-10) and long test (11-20) conditions, respectively. These data reveal that Method B constantly yields smaller standard error estimates overall across the theta scale. The difference is most pronounced at moderate to high correlation conditions and for the long test conditions.

The random error was observed to decrease as the test length increased regardless of the information computation method used. The SEs showed that random measurement error was greater with Method A than Method B for all conditions.

The SEs of the short-test moderate-discrimination conditions (conditions 1-6) were smaller with Method B than with Method A across the theta scale. The SEs with Method B became smaller as the correlation between math and science traits increased. The SEs of the short-test high-discrimination conditions (conditions 6-10) were also smaller with Method B, and became smaller as the correlation between math and science traits increased. However, Method B SEs were relatively greater than Method B SEs at the upper and lower ends of the theta scale.

A similar pattern was observed under long test conditions (conditions 11-20) with an increased discrepancy between Method A and Method B SEs. Results indicate that the decrease observed in the SEs for the moderate discrimination test conditions is almost constant across the theta scale for both short and long test conditions. With the high discrimination conditions, results show that the increases in item discrimination are associated with decreases in SEs. Furthermore, Method A SEs were relatively smaller than those of the Method B at the ends of the theta scale, and the Method A SEs, and were greater than those of Method B at the mid theta scale.



1. Moderate discrimination/0.1 Correlation

2. Moderate discrimination/0.3 Correlation

3. Moderate discrimination/0.5 Correlation

4. Moderate discrimination/0.7 Correlation

5. Moderate discrimination/0.9 Correlation

6. High discrimination/0.1 Correlation

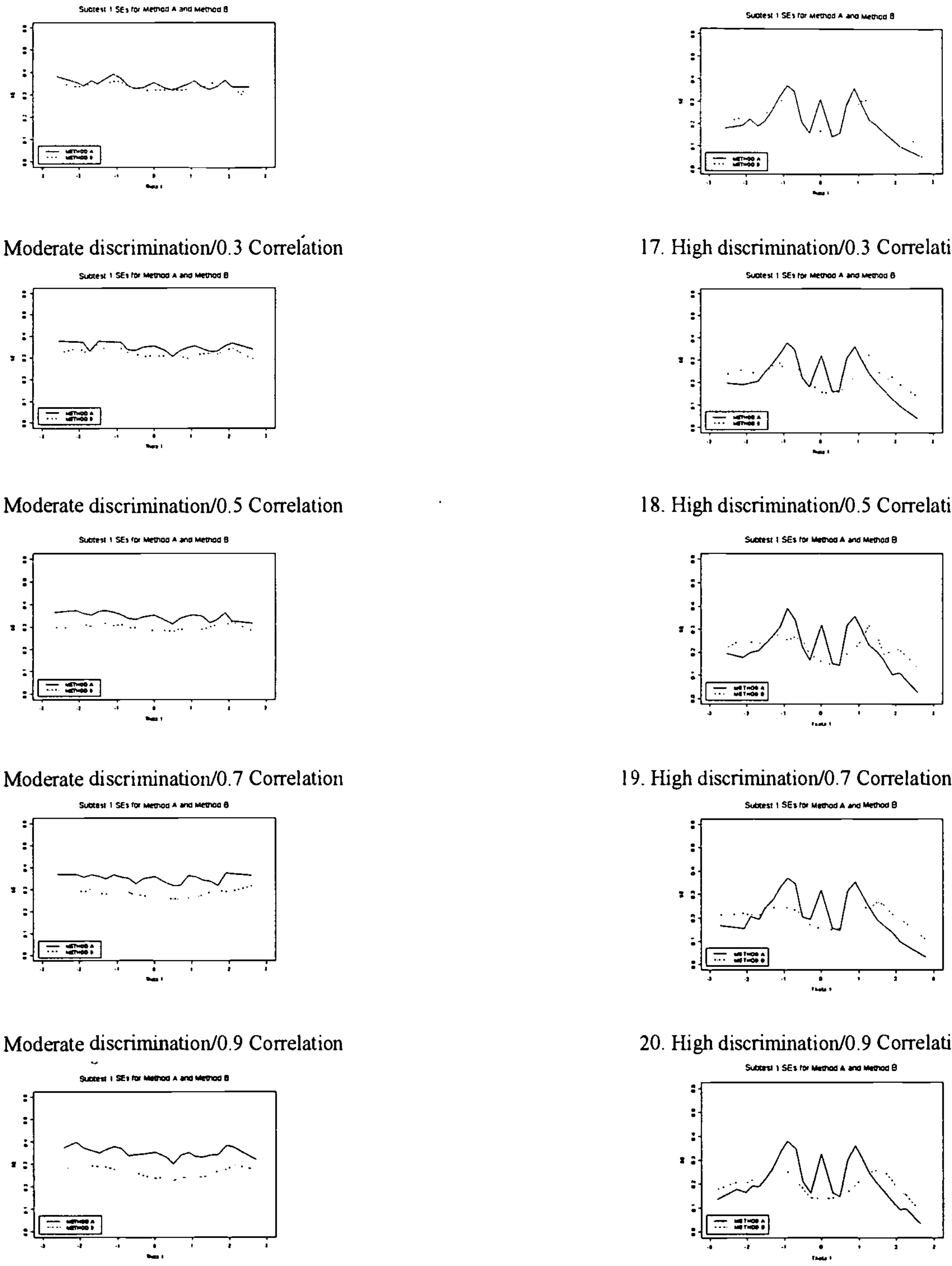
7. High discrimination/0.3 Correlation

8. High discrimination/0.5 Correlation

9. High discrimination/0.7 Correlation

10. High discrimination/0.9 Correlation

Figure 6. The Standard Error of the Math Ability Estimates for the Short Test Conditions



11. Moderate discrimination/0.1 Correlation

12. Moderate discrimination/0.3 Correlation

13. Moderate discrimination/0.5 Correlation

14. Moderate discrimination/0.7 Correlation

15. Moderate discrimination/0.9 Correlation

16. High discrimination/0.1 Correlation

17. High discrimination/0.3 Correlation

18. High discrimination/0.5 Correlation

19. High discrimination/0.7 Correlation

20. High discrimination/0.9 Correlation

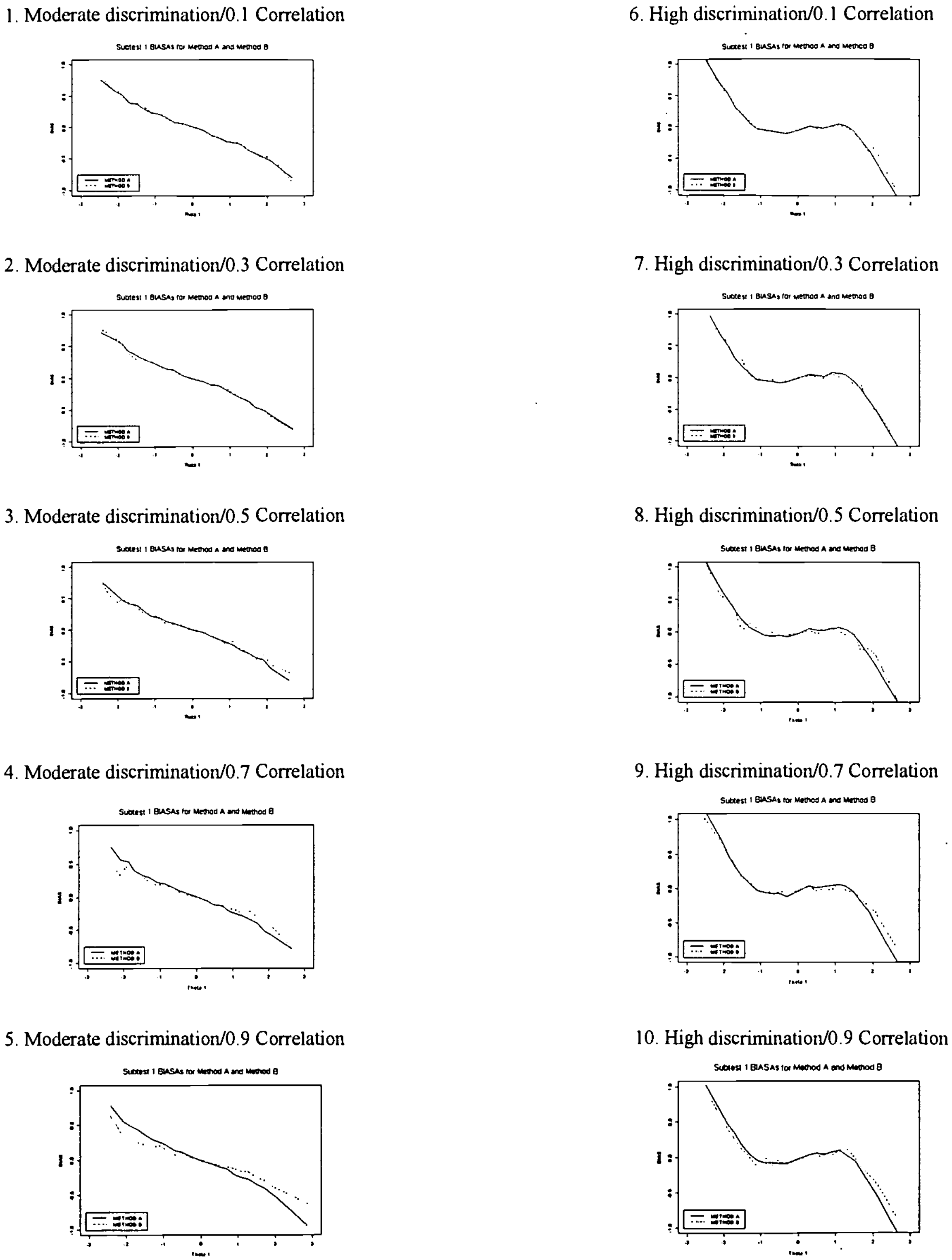
Figure 7. The Standard Error of the Math Ability Estimates for the for the Long Test Conditions

Bias

Figure 8 and 9 show the conditional bias plots for the 10 short test (conditions 1-10) and 10 long test conditions (conditions 11-20) for the math test, respectively. These plots indicate that the estimated bias did not show large discrepancies between the two Methods. The bias estimates were relatively smaller (a) for the long test conditions than those of the short test conditions, and (b) for the high discrimination conditions than those of the moderate discrimination conditions. This difference became larger as the correlation level was increased.

The bias under the moderate-discrimination short-test condition was smaller with Method B than it was with Method A when trait correlation was 0.7 and even smaller when trait correlation was 0.9. This pattern was also emerged in the high discrimination-short test conditions, however, lesser in degree. In the long test-moderate discrimination conditions, bias of Method B was slightly greater for extreme ability levels, and was slightly smaller for middle ability levels. In the long test-high discrimination conditions, bias of Method B was greater for extreme ability levels, and was smaller for middle ability levels. The two methods yielded similar estimated bias when the correlation between the two traits was 0.9. In the long-test high-discrimination condition, the bias was smaller with Method B, and Method B performed better as the trait correlation increased. The bias of the 0.1 correlation-long-test-high discrimination condition, condition 16, was almost identical for both methods.

Overall, Method B bias became smaller when the correlation between math and science traits was moderate to high, test items were highly discriminating, and test length was longer, in this order. Method B bias was found to be slightly smaller than that of Method B when the trait correlation increased as the ability level became extreme under all conditions.


1. Moderate discrimination/0.1 Correlation

2. Moderate discrimination/0.3 Correlation

3. Moderate discrimination/0.5 Correlation

4. Moderate discrimination/0.7 Correlation

5. Moderate discrimination/0.9 Correlation

6. High discrimination/0.1 Correlation

7. High discrimination/0.3 Correlation

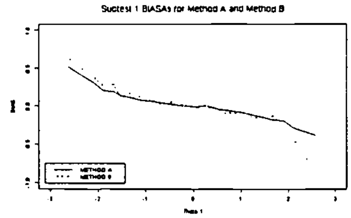
8. High discrimination/0.5 Correlation

9. High discrimination/0.7 Correlation

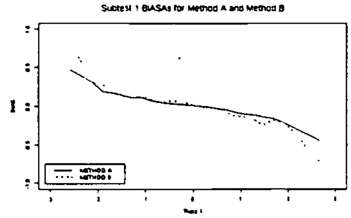
10. High discrimination/0.9 Correlation

Figure 8. Estimated Bias of the Math Ability Estimates for the Short Test Conditions

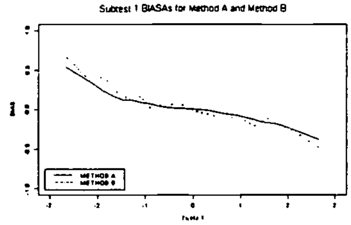
11. Moderate discrimination/0.1 Correlation



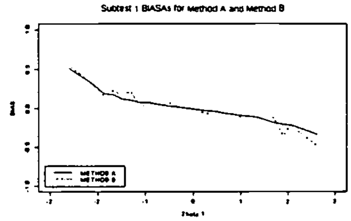
12. Moderate discrimination/0.3 Correlation



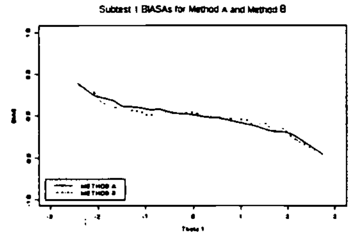
13. Moderate discrimination/0.5 Correlation



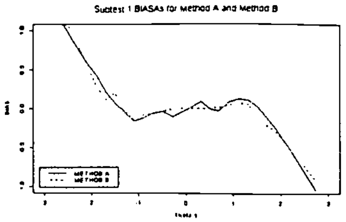
14. Moderate discrimination/0.7 Correlation



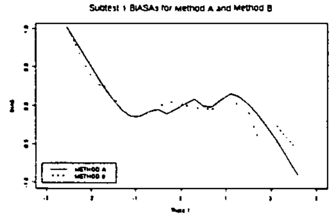
15. Moderate discrimination/0.9 Correlation



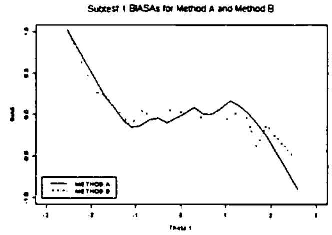
16. High discrimination/0.1 Correlation



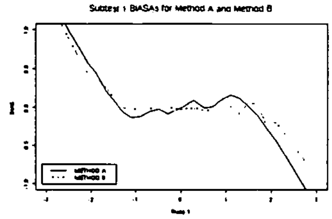
17. High discrimination/0.3 Correlation



18. High discrimination/0.5 Correlation



19. High discrimination/0.7 Correlation



20. High discrimination/0.9 Correlation

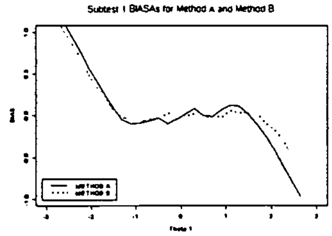


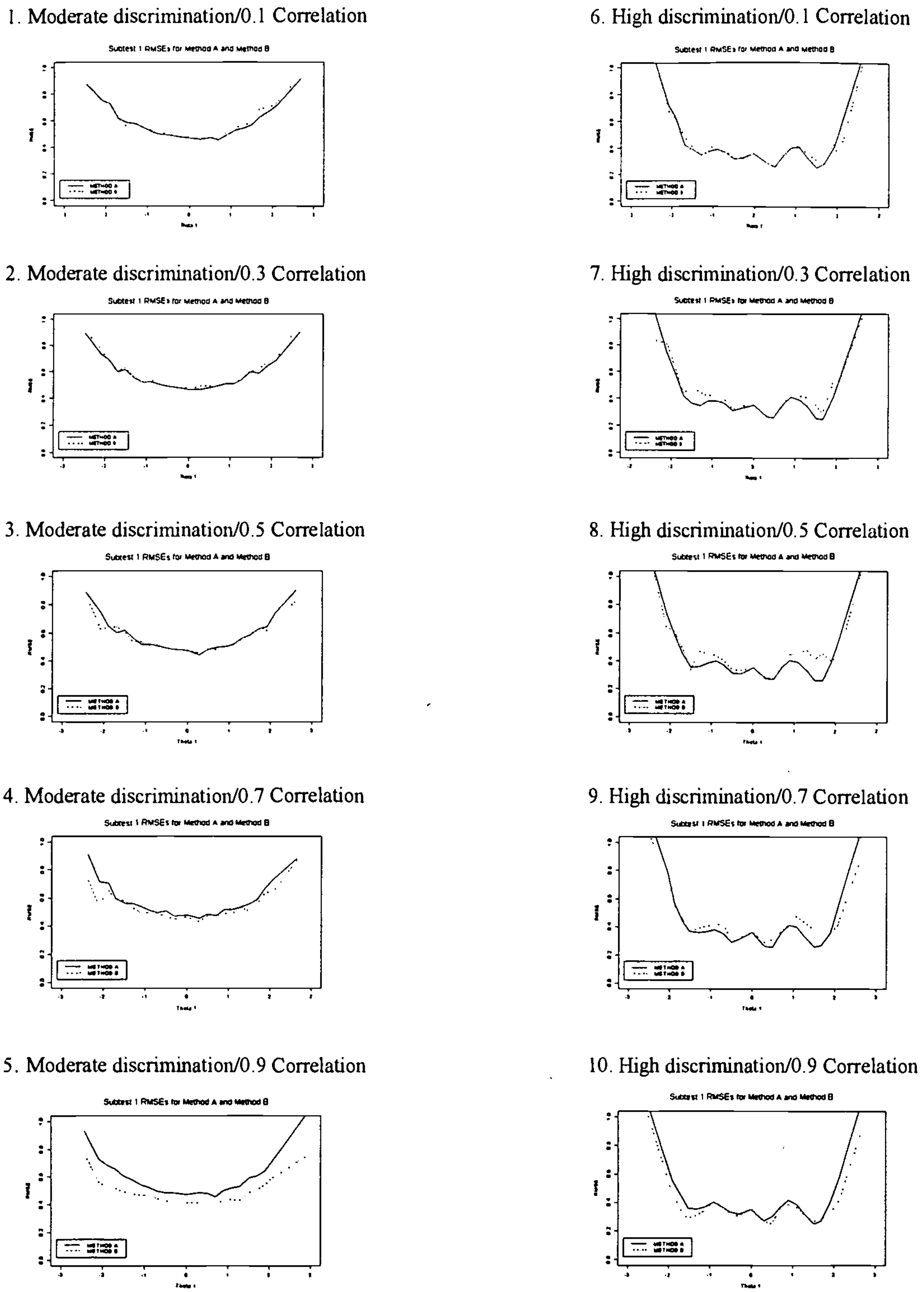
Figure 9. Estimated Bias of the Math Ability Estimates for the Long Test Conditions

Somewhat contradictory to what was expected, ability estimates with Method B were only slightly more biased than those with Method A in general. One possible explanation is that the using the EAP procedure with both Method A and Method B, instead of the ML, procedure led to biased estimates for both methods.

Root Mean Square Error (RMSE)

As noted earlier, RMSE is a function of both SE and bias ($RMSE^2=SE^2+bias^2$). Consequently, the results for RMSE are related to those already discussed for SE and bias. Figure 10 and 11 show the conditional RMSEs for math ability at short-test conditions (conditions 1-10) and long-test conditions (conditions 11-20), respectively. Short-test condition RMSEs were found to be higher than those of the long-test conditions in general. The RMSEs increased as the ability level of the examinee became extreme at both short and long-test conditions, and this increase was greater at the high-discrimination conditions. This could be due to observed higher bias in the high-discrimination conditions for extreme abilities.

The results indicate that the RMSEs of the short test-moderate discrimination conditions were smaller with Method B than those with Method A as the correlation level increased. The RMSEs of the Method B were the smallest when compared to those of Method A for condition 5, where the math and the science trait correlation was 0.9. The RMSEs of the short test-high discrimination conditions were slightly greater with Method B than with Method B for 0.3, 0.5, and 0.7 correlation conditions.



1. Moderate discrimination/0.1 Correlation

6. High discrimination/0.1 Correlation

2. Moderate discrimination/0.3 Correlation

7. High discrimination/0.3 Correlation

3. Moderate discrimination/0.5 Correlation

8. High discrimination/0.5 Correlation

4. Moderate discrimination/0.7 Correlation

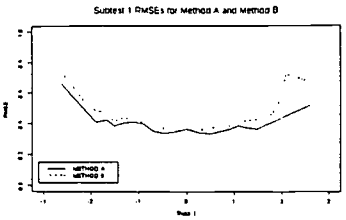
9. High discrimination/0.7 Correlation

5. Moderate discrimination/0.9 Correlation

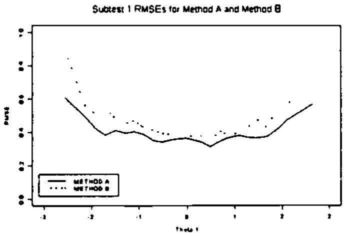
10. High discrimination/0.9 Correlation

Figure 10. Estimated Root Mean Square Error (RMSE) of the Math Ability Estimates for the Short Test Conditions

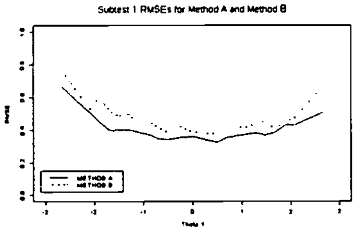
11. Moderate discrimination/0.1 Correlation



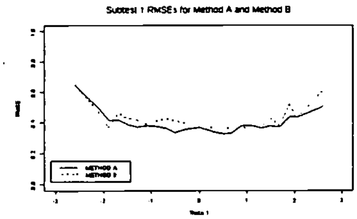
12. Moderate discrimination/0.3 Correlation



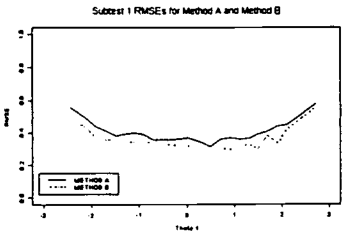
13. Moderate discrimination/0.5 Correlation



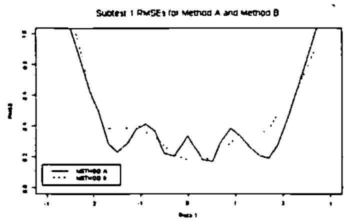
14. Moderate discrimination/0.7 Correlation



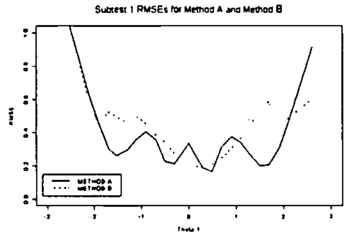
15. Moderate discrimination/0.9 Correlation



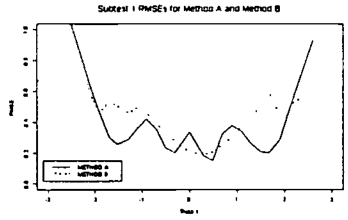
16. High discrimination/0.1 Correlation



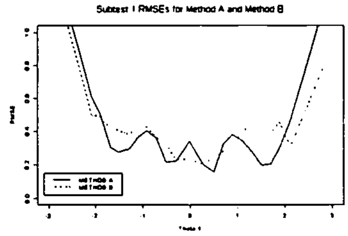
17. High discrimination/0.3 Correlation



18. High discrimination/0.5 Correlation



19. High discrimination/0.7 Correlation



20. High discrimination/0.9 Correlation

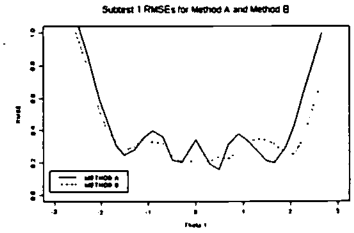


Figure 11. Estimated Root Mean Square Error (RMSE) of the Math Ability Estimates for the Long Test Conditions

In the short test-high discrimination conditions, the RMSEs were greater with Method B, and for extreme ability levels in general. However, when compared, Method B RMSEs were smaller than Method A RMSEs for extreme ability levels. The only short test-high discrimination condition with overall smaller Method B RMSEs was condition 10, where the math and the science trait correlation was 0.9.

The RMSEs of the long test conditions (conditions 11-20) were greater for Method B in general. Under the long test conditions, RMSEs were greater for Method B for 0.1, 0.3, and 0.5 correlation conditions for both moderate and high discrimination conditions. The RMSEs were approximately the same for Method A and Method B for long-moderate and high discrimination conditions when the trait correlation was 0.7. The RMSEs were smaller with Method B for 0.9 correlation condition only for both moderate and high discrimination conditions. The only short test-high discrimination condition that resulted with overall smaller RMSEs for Method B was condition 10, where the math and the science trait correlation was 0.9.

The RMSEs of the long test conditions (conditions 11-20) were greater for Method B in general. Under the long test conditions, RMSEs were greater for Method B for 0.1, 0.3, and 0.5 correlation conditions for both moderate and high discrimination conditions. The RMSEs were smaller with Method B for 0.9 correlation condition only for both moderate and high discrimination conditions.

Summary

The finding for SE indicate that the random measurement error is more evident with the Method A, the in-scale information method, than with the Method B. This SEs difference were more prevalent conditions across the theta scale for the short test conditions. The only conditions were high discrimination-moderate correlation conditions where SEs with Method B were greater than those with Method A only for extreme ability levels.

The results for bias show that ability estimates at high and low theta levels are pulled toward the mean (pulled inward) with both methods. This inward bias was expected since the EAP Bayesian procedure was used in obtaining ability estimates. The results indicate that Method A and Method B bias did not differ for the short test conditions with the only exception of 0.9 correlation conditions. The biases were smaller for Method B for moderate and high discrimination conditions with 0.9 trait correlation (conditions 5 and 10). Method B bias also did not differ from the Method A bias for long test-moderate discrimination conditions (conditions 11-15). The bias associated with Method B ability estimates were smaller for the extreme ability levels and greater for middle range ability levels for long test-high discrimination conditions, more so when the correlation between the two traits increased.

The RMSEs reflecting the combined effects of SEs and Bias were smaller with Method B for moderate discrimination conditions only when the math and science trait correlation was 0.7 and became smaller when the math and science trait correlation was 0.9 for both short and long test conditions. However, the Method B RMSEs of the high discrimination conditions were greater with 0.1, 0.3, and 0.5. The discrepancy almost disappeared as the trait correlation was increased to 0.7. At 0.9 correlation level, Method B RMSEs were smaller than those of Method A. In sum, short test Method B RMSEs were very similar to those of Method A, and become smaller than Method A RMSEs as the correlation level increased from 0.5 to 0.7, and 0.7 to 0.9.

In the short test conditions, Method B RMSEs were smaller than those of Method A for the moderate discrimination condition. In the long-test conditions, Method B RMSEs were greater than those of Method A for moderate discrimination conditions. The discrepancy dissapeared as the correlation level increased to 0.7, and the RMSEs of Method B became smaller as the correlation level increased to 0.9. The Method B RMSEs were approximately the same with the Method A RMSEs in the long test-high discrimination conditions. However, with increasing correlation level Method B RMSEs were smaller than Method A RMSEs. This was most pronounced with 0.9 correlation level.

The performance of Method A and Method B RMSEs in the long test-high discrimination conditions was different across the theta scale. Method B RMSEs were smaller than Method A RMSEs for the mid ability range, and were smaller than Method A RMSEs for the extreme ability range. The decrease in Method B RMSEs relative to the Method A RMSEs was also observed at the extreme ability range. That is, as the correlation level increased to 0.9 Method B not only performed better than Method A in the mid ability range but also at the extreme ability range. The RMSEs computed for the long test conditions required higher trait correlation to recover true ability parameters, than those for the short test conditions. That is, when short and long test moderate discrimination conditions are compared, short RMSEs were smaller with Method B.

Discussion

One of the most important tasks facing measurement specialists is an increasing demand from stakeholders that large-scale assessments should provide more test-scores that can be used for diagnostic purposes. However, it is rare to have enough items in a test battery to report content specific scores within subtests with an acceptable level of reliability. Increasing the number of items per sub-content domain is not an attractive solution for both test-developers and test-takers. An alternative is to investigate procedures that would make use of already available information in the response data.

This study was conducted to evaluate the expected increase in the reliabilities of subscale scores when all items composing a test battery are allowed to contribute the subscale score to the extent that they retain discrimination power with respect to the trait intended to be measured by these subscales.

The choice between the in-scale, Method A, and in- and out-of-scale, Method B, procedures would seem to depend primarily on SE and bias. The SEs were found to be constantly smaller with the in- and out-of-scale information method than with the in-scale information method, favoring the out of-scale information method. If the purpose is to rank order examinees according the composite trait measured, bias will be lesser of concern than SE or random error, since bias will not effect the rank order of examinees with respect to a trait as long as the bias is monotonically related to the ability estimates. Wang and Vispoel (1998) listed three situations where bias is more important than SE as in comparing group means, in referencing ability estimates from different tests, and making mastery/non-mastery or other classification decisions. In such cases Method B

It was shown that the in- and out-of-scale information procedure increased the reliability of the subscale scores without increasing the number of test items when traits measured are moderately or higly correlated. The out-of-scale information method performed better (smaller SEs and RMSEs) with the short test-moderate discrimination conditions when compared to the in-scale information method as the correlation between the traits measured increased. This finding is likely to make the alternative information computation method more attractive to the users because tests in practice are likely to be comprised of moderately discriminating items and the domain scales that need precision increases are usually short tests, i.e., tests with less than 20 items. Using out-of-scale information is likely to produce reliable subscores that can be used for diagnostic purposes, where examinees performance repertoire can be decried in terms of what the examinee can and can not do.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT Calibration of compensatory and non-compensatory multidimensional items. Applied Psychological Measurement, 13(2), 113-127.

Ackerman, T. A. (1996). Graphical representations of multidimensional item response theory. Applied Psychological Measurement, 20, 311-329.

Ackerman, T. A., & Davey, T. C. (1991). Concurrent adaptive measurement of multiple abilities. Paper presented at the annual meeting of the American Educational Research Association (Chicago, IL, April).

Ansley, R. A., & Foryth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Davey, T. C., & Hirsh, T. M. (1991). Examinee discrimination as measurement Properties of multidimensional tests. Paper presented at the annual meeting of the National Council on Measurement in Education (Chicago, IL, April).

Davey, T., & Spray, J. (1999). PIC: Pretest item calibration computer program. {Computer Program}. Iowa City IA: ACT Inc.

Gifford, J. A., and Swaminathan, H. (1990). Bias and the Effect of Priors in Bayesian Estimation of Parameters of Item Response Models. Applied Psychological Measurement, 14(1), 33-43.

Hambleton, R. K. and Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston, MA: Kluwer. Nijhoff Publishing.

Junker, B. W & Stout, F. W, (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In Laveault, D., Zumbo, B. D., Gessaroli, M. E., & Boss, M. W. (Eds.), Modern theories of measurement: Problems and Issues (pp. 51-84). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. Applied Psychological Measurement, 16(3), 279-293.

Miller, T. R., & Hirsh, T. M. (1990). Cluster analysis of angular data in applications of multidimensional item response theory. Paper presented at the annual meeting of the Psychometric Society, Princeton NJ, June.

Miller, T. R., & Hirsh, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. Applied Psychological Measurement, 5(3), 193-211.

Misleyv, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Moorisville IN: Scientific Software.

Muraki, E. (1992). RESGEN: Item Response Generator. Research Report (RR-92-7), Educational Testing Service, Princeton, New Jersey.

Nandakumar, R. (1991). Traditional versus essential dimensionality. Journal of Educational Measurement, 28(2), 361-373.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. Applied Psychological Measurement, 9(4), 401-412.

Reckase, M. D. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement, 15(4), 361-373.

Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, 25(3), 193-203.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimates. Psychometrika, v. 55, 293-325.

Wainer, H., Sheehan, M. Wang, X. (2000). Some paths toward making praxis scores more useful. Journal of Educational Measurement, 37(2), 113-140.

Wang, M. (1986). Fitting a unidimensional model to multidimensional response data. Paper presented at the ONR Contractors Conference (Gatlinburg, TN April).

Wang, M. (1988). Measurement bias in the application of a unidimensional model to multidimensional item response data. Paper presented at the annual meeting of the National Council on Measurement in Education (New Orleans, LA April).

Wang, T., & Vispoel, W. P. (1999). Properties of ability estimation methods in computerized adaptive testing. Journal of Educational Measurement, 35(2), 109-135.

Yen, W. M. (1987). A Bayesian/IRT index of objective performance. Paper presented at the annual meeting of the Psychometric Society (Montreal, Quebec, Canada June).

Zhang, J., & Wang, M. (1998). Relating reported scores to latent traits in a multidimensional test. Paper presented at the annual meeting of the American Educational Research Association (San Diego, CA April).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Analyzing Multidimensional Response Data Structure Represented by Unidimensional models to Increase the Precision of Scoring Using Out-of-Scale Information	
Author(s): Nilufer K. Capar	
Corporate Source: Nilufer K. Capar	Publication Date: November 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
↑ ☒	↑ □	↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: nilufer K. Capar	Printed Name/Position/Title: Dr. Nilufer Kahraman Capar	
Organization/Address: 603 Fulton Rd. E-45 Tallahassee, FL 32312	Telephone: 850-3867578	Fax:
	E-mail Address: nkk3877@ garnet.acns.fsu.edu	Date: 8/7/2001

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM: